December 4, 2024

***E-Filed***

The Honorable Thomas S. Hixson
United States District Court for the Northern District of California
San Francisco Courthouse, Courtroom E – 15th Floor
450 Golden Gate Avenue
San Francisco, CA 94102

Re: *Kadrey et al. v. Meta Platforms, Inc.*; Case No. 3:23-cv-03417-VC-TSH

Dear Magistrate Judge Hixson:

Plaintiffs in the above-captioned action ("Plaintiffs") and Defendant Meta Platforms, Inc. ("Meta") jointly submit this letter brief regarding issues related to Plaintiffs' RFP Nos. 118 and 119. The parties met and conferred on November 20, November 27, and December 2, 2024 but were unable to reach a resolution on the following issues.

# I. Plaintiffs' Statement

Plaintiffs are entitled to information about how Meta obtained copyrighted works, including Plaintiffs' works; the number of copies it made; and how it used those works in training and operationalizing its Llama models. This includes information about Meta's alteration and distribution of copyrighted works, Llama's propensity to regurgitate copyrighted material, and the features that Meta built into Llama models to mitigate infringing outputs. RFPs 118-119 are intended to answer these questions, but Meta continues to withhold relevant information.

## A. RFP No. 118: Source Code and Post-Training Data

RFP 118 seeks materials related to Meta's efforts to prevent Llama from emitting copyrighted material. Ex. A. Meta provided some responsive material via documents and testimony, including in the so-called "pre-training" phase. But a significant hole remains: Meta's *post*-training materials are largely missing, including datasets used to teach Llama models *not* to emit copyrighted data, and the source code Meta addressing "memorization" and "regurgitation" issues.

This missing data exists—Meta acknowledges a so-called mitigation strategy covering copyright-protected materials and other content. Meta witnesses also have explained that Meta likely monitors Llama output on a real-time basis, as it apparently is cheaper to remove regurgitated material once it appears than to prevent it from appearing in the first place. *See* Ex. B. All of this is relevant, as is the general topic of how the models memorize and regurgitate. For example, if a model can output or regurgitate copyrighted material on which it has trained, that is direct evidence of copying. It is also probative of the form in which the models functionally store a copy, or "memory," of training data. The information Meta has disclosed affirmatively—including testimony from nearly every relevant witness—makes clear that memorization is a recurring and endemic problem with Llama. Meta's efforts in connection with handling the limited reality of what its models do—among other things, memorize and regurgitate—are thus relevant to Plaintiffs' core claim and Meta's defenses, including Meta's argument about purported transformative use, i.e., that material on which Llama trains is "transformed" when Llama creates outputs. The post-training data that Meta uses to regulate Llama's outputs bears on this analysis.

As a compromise, Plaintiffs requested two specific post-training datasets from Meta: (i) the data mentioned in Sections 3 and 4.2 of Meta's Llama 2 Paper,[1] and (ii) the data mentioned in Sections 4.2 and 5.4.3 of Meta's Llama 3 Paper.[2] The "Llama Papers" are publicly available research papers published by Meta. The cited sections describe Llama's mitigation protocols and repeatedly reference post-training data Meta refuses to produce here. The Court should order the production of this data.

Relatedly, the source code that Meta *has* produced is likely deficient. Meta's Source Code Room contains an unorganized mix of scripts and experimental code, not the organized set of code needed to support a sophisticated LLM. If it would assist the Court, Plaintiffs are prepared to submit an expert declaration from Crista Lopes, PhD regarding the jumbled state of Meta's Source

---

1 https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

2 https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

Code Room. Dr. Lopes has personally attended the Source Code Room and has observed its contents.

A party must produce data as it is kept in the usual course of business or must organize and label it to correspond to the categories in the request. Fed. R. Civ. P. 34(b)(2)(E)(i); *see*, *e.g.*, *Davis v. Pinterest, Inc.*, 2021 WL 3045878, at *6 (N.D. Cal. July 20, 2021) (Hixson, J.) (confirming that this rule applies to ESI not just hard-copy documents). This means ESI, including code, must be produced "in whatever organizational structure [it is] stored in and not just randomly thrown together . . ." *Id.* at *7. What Meta has done is provide Plaintiffs with a messy sandbox of source code instead of the actual, comprehensive, well-organized code that underlies the Llama models as Meta itself uses it. The Court should order Meta to remedy the deficiency and immediately make the code and configuration files available in the organized format Meta actually uses for its Llama models.

**B. RFP No. 119: Treatment of Copyrighted Material**

RFP 119 involves the processing of copyrighted material used to train the Llama models, including Meta's storage and deletion of copyrighted material. Ex. A. To start, Meta refuses to produce or identify *all* copies it made of copyrighted works, including Plaintiffs' works. This discovery bears on the breadth of use of the copyright protected material at issue—if Meta repeatedly downloaded Plaintiffs' works without their consent and/or distributed them to others, each act is a separate instance of infringement. Unlike other RFPs that this Court held did not encompass "all" copies, *see* Dkt. No. 288 at 5, RFP 119 expressly does. Meta should produce the copies or catalogue them.

Further, Plaintiffs recently learned that Meta stripped copyright management information ("CMI") from copyrighted works that it torrented from third-party databases to train certain Llama models. *See* Dkt. No. 301-2 at 3-6.[3] As part of that torrenting, Meta also *uploaded* those works to third-party databases—effectively turning Meta into a distributor of pirated copyrighted material. *Id.* Yet Meta has not searched for or produced the corpus of documents in its possession relating to its CMI stripping and torrenting of copyrighted material, including its decision *not* to strip CMI from unprotected works. These actions fall within RFP 119's request for information regarding "storage and deletion of copyrighted material." Plaintiffs therefore request an order requiring Meta to (1) produce or catalogue all copies it made of copyrighted material, and (2) search the custodial files of its 15 custodians, plus relevant non-custodial databases (including work email, Workplace, Workplace Chat, and WhatsApp),[4] for documents and communications regarding Llama and (a) stripping or removal of CMI from literary works, or (b) torrenting" of data that includes literary works.

---

[3] Plaintiffs' Motion for Leave and Proposed Third Amended Complaint (Dkt. 301) both outline existing evidence of Meta's CMI stripping and torrenting. Regardless of whether that Motion is granted, the evidence still bears on willful copyright infringement, which is independently relevant to Plaintiffs' claims.

[4] Meta recently confirmed for the first time on a December 2 meet and confer that Meta's work email and Workplace data including Workplace Chat are maintained on centralized systems. Thus, these are "non-custodial" data sources that should be searched for all relevant hits, not just hits involving the 15 identified custodians.

## II. META'S STATEMENT

Discovery in this matter closes in seven business days. Rather than focusing on taking the remaining depositions[5], Plaintiffs continue to seek to compel Meta to produce documents far outside the relevant issues in this case, that impose a disproportionate burden, and relitigate settled discovery issues Meta has addressed a reasonable scope for RFPs 118-119. The Court should deny relief.

**Meta Has Adequately Responded to RFP 118.** Plaintiffs' RFP 118 seeks documents "relating to any efforts, attempts, or measures implemented by Meta to prevent Llama Models from emitting or outputting copyrighted material." However, Judge Chhabria dismissed Plaintiffs' claims that the *output* of the Llama models constituted copyright infringement by Meta. ECF 56 at 1-3. Thus, RFP 118 is of ancillary relevance, at most, to the only claim remaining in this case: whether Meta's use of datasets allegedly containing copies of Plaintiffs' books to train the Llama models constitutes copyright infringement. ECF 56 at 1.

Despite the tangential nature of RFP 118, Meta has produced numerous responsive documents, as Plaintiffs acknowledge. In addition, in an effort to avoid unnecessary motion practice, Meta also has made available to Plaintiffs (1) the code for its products incorporating the Llama models, including Meta AI, as well as (2) any corresponding code related to post-output filtering (thereby addressing an issue that Plaintiffs have highlighted above). Meta also is collecting and will shortly produce additional documents concerning output mitigations specific to the Meta AI platform.

Notwithstanding Meta's adequate and proportional response to RFP 118, Plaintiffs demand more and seek documents that are largely unrelated to RFP 118. It was no "compromise" for Plaintiffs to seek the alleged "data" mentioned in the cited sections of the Llama 2 and 3 papers. Together, these sections contain over 15 pages of dense, technical text describing a variety of post-training *processes* and *techniques*, not any discrete "data" that can be readily collected. Tracking down any "data" related to these sections of the paper would necessitate, among other things, additional witness interviews to understand and identify what data could be responsive, followed by searches, collection, and review–all before the close of discovery. Furthermore, these sections of the Llama papers sweep in matters that have little to do with "measures implemented by Meta to prevent Llama Models from emitting or outputting copyrighted material" (the subject of RFP 118) such as output helpfulness and safety-related issues (e.g., toxicity and bias). And even if any of the cited sections concerned copyright-related mitigations, Plaintiffs' theory of relevance fails. Plaintiffs suggest that if a model can memorize training data, that is "direct evidence of copying," but "post training data"—which, by definition, is used *after* the training that is the subject of Plaintiffs' lone remaining claim—will not address whether Plaintiffs' books were used for or memorized during training. Plaintiffs' suppositions should be addressed by (1) the datasets that they allege contain their books, which have been produced by Meta, and (2) the Llama models themselves, which are publicly available. Plaintiffs also baldly assert that "post-training data… bears on" the issue of transformativeness, but they do not explain how it is connected to the issues in dispute.

---

[5] In the last two weeks of discovery, Meta is producing 8 witnesses for deposition and is taking at least 3 depositions of Plaintiffs. *See* ECF No. 302.

Finally, for the first time in their letter brief, Plaintiffs vaguely allege that "Meta's Source Code Room contains an unorganized mix of scripts and experimental code." Plaintiffs' allegation is simply wrong. Meta provided source code in the manner it is kept in the ordinary course of business and as governed by the Stipulated Protective Order. Notably, the source code has been available to Plaintiffs since May, yet they waited until this brief to assert unspecified issues with the source code's organization. In any case, Plaintiffs' complaints about the source code have *nothing to do with* RFP 118. To the extent these baseless complaints are timely (which Meta disputes), the allegations are not properly presented here; Plaintiffs did not raise this issue in any of their correspondence with Meta or in either of the two meet and confers concerning RFP 118.

**Meta Has Adequately Responded to RFP 119.** Plaintiffs' massive demands regarding RFP 119—a catalog of every single copy made by Meta of any dataset potentially containing copyrighted works, and a Meta-wide search for all documents and communications regarding Llama and CMI or torrenting—overlook what RFP 119 actually seeks and are a poorly disguised effort to relitigate this Court's prior decision rejecting Plaintiffs' overly burdensome demands for all copies of training data at Meta that may contain Plaintiffs' works. ECF 288. It does not request *any* copies of training data, but rather "All Documents and Communications, including source code, relating to the *processing* of copyrighted material used in training Llama Models, including storage and deletion of copyrighted material." (emphasis added)

And even if RFP 119 could be construed with such breadth, it would be overbroad, unduly burdensome, and disproportionate to Plaintiffs' needs. Plaintiffs' demand that Meta "produce or catalogue all copies it made of copyrighted material," already once rejected by the Court, *see id.*, is an impossible and extraordinarily burdensome undertaking, particularly with seven business days left in discovery. "[C]opyrighted material" is broad and undefined and "copies" could include all or portions of datasets that contain many terabytes of data. Nor is it relevant: whether Meta made one copy of one or multiple copies is not pertinent, as a copyright plaintiff is entitled to at most one recovery per work, regardless of the number of copies made. *See Desire, LLC v. Manna Textiles, Inc.*, 986 F.3d 1253, 1266 (9th Cir. 2021) (explaining that the intent of the Copyright Act is "to constrain the award of statutory damages to a single award per work, rather than allowing a multiplication of damages based on the number of infringements" (quotations omitted)).

Plaintiffs also seek to shoehorn their dismissed CMI claim into RFP 119. This claim was dismissed over a year ago, *see* ECF 56 at 3, and thus is an improper basis for an RFP.[6] Plaintiffs also seek to shoehorn their "torrenting" allegations into RFP 119, but that subject matter has nothing to do with this RFP and instead is the subject of another RFP (No. 85) not at issue in this

[6] After the close of business on Thanksgiving Eve, and without any prior notice to Meta, Plaintiffs moved for leave to file an amended complaint, including to revive their previously dismissed CMI claim. ECF 301, Appx. A, at 17-18. Meta will respond to Plaintiffs' motion in due course. For now, Meta observes that (1) there is no CMI claim presently in the case and hasn't for over a year and (2) Plaintiffs' demands for more CMI-related discovery are directly contrary to their argument in the Motion for Leave that the amendment would not require any "additional discovery beyond what is needed for the copyright infringement claim." ECF 301 at 1. Moreover, Plaintiffs' assertion that they only "recently learned that Meta stripped [CMI]" is betrayed by their reliance on a document produced by Meta in May (ECF 301-7) addressed with Meta witnesses in September depositions.

letter brief that Meta is producing documents for. Plaintiffs have not explained or substantiated how their sweeping demands under RFP 119 are relevant to willfulness, and they are not.[7]

Finally, Plaintiffs again re-raise demands that Meta conduct company-wide searches of employee emails, chats and WhatsApp messages in response to RFP 119.[8] This demand to rewrite the ESI order was rejected by the Court just a few weeks ago, noting it "has no merit now." ECF 279 at 3-4. As Meta previously explained, Plaintiffs' demands would erase the limitations on custodians in the Stipulated ESI Order and would mark a fundamental change on the scope of discovery with only 7 business days left. It would also impose enormous burdens on Meta, as providing custodial data for just a few custodians requires many weeks of work, as addressed in previous briefing on requests for a smaller number of custodial data searches. ECF 196 at 2 (recognizing that "[a]dding five document custodians is a big change, not a small change, to the scope of Meta's document production obligations in this case").

## III. PLAINTIFFS' REPLY

**RFP 118** is not "of ancillary relevance." Llama's tendency to "memorize" training data is additional evidence of copyright infringement even beyond the copying done at other stages of Meta's work. And Meta's efforts to mitigate/conceal Llama's emission of copyrighted material shows willfulness. Meta is also wrong that Plaintiffs never raised the issue of its source code organization. Plaintiffs have done so repeatedly. Each time, Meta represented there were no issues to address or that it would address them. *See, e.g.*, Dkt. No. 199 at 4-5 n.4 (Meta arguing on 10/3 it "retain[ed] the file structure and organization of the repositories used at Meta"). In any event, Meta doesn't dispute the code it's made available—including ***four terabytes*** just yesterday—must be provided as it's kept in the usual course of business (i.e., in the way Meta's employees use it). Accordingly, the Court should grant the motion as to this request.

Contrary to Meta's argument that **RFP 119** does not encompass copies of training data, the RFP defines "processing" to *include* "storage" of copyrighted material. In citing burden, Meta fails to tell the Court it *already* knows where it stores copies, namely in storage locations called Manifold, GTT, and Hive. To the extent Meta argues that even just identifying these copies is "an impossible and extraordinarily burdensome undertaking" (perhaps because there are many), Meta never raised *this* burden objection in the M&C, its discovery response, or other briefing (*see* Dkt. 267 at 31). Instead, it just cited burden with other boilerplate objections. If Meta won't produce the actual copies, the Court should order Meta to provide that information in a declaration instead.[9]

---

[7] Plaintiffs will likely claim that Meta's purported "removal" of copyright notices from datasets is evidence that Meta tried to conceal infringement, but this theory is contrary to the documents and testimony in the case and also makes no sense. Deleting copyright notices from training data cannot conceal Meta's use of that data, because it is not otherwise publicly disclosed.

[8] Nearly identical issues were addressed in the November 9, 2024 omnibus letter brief addressing Plaintiff's late-raised disputes on Existing Written Discovery (ECF 267, Issue #1), which the court denied. (ECF 288.) Nevertheless, Plaintiffs' serial re-litigation of discovery matters is not stopping, as Plaintiffs have indicated they will serve Meta with yet another letter brief seeking additional broad custodial searches of email, chat and WhatsApp messages in response to other RFPs. Meta expects that Plaintiffs will file that letter brief with the Court soon.

[9] Meta argues the Court should deny Plaintiffs' requests because Meta can't comply before 12/13. But that is not a reason to deny Plaintiffs the discovery to which they're otherwise entitled. Rather, the Court should grant the motion and if Meta needs additional time to comply then the parties can ask Judge Chhabria.

By: /s/ *Bobby Ghajar*

Bobby A. Ghajar
Colette A. Ghazarian
**COOLEY LLP**
1333 2nd Street, Suite 400
Santa Monica, CA 90401
Telephone: (310) 883-6400
Facsimile: (310) 883-6500
Email: bghajar@cooley.com
cghazarian@cooley.com

Mark R. Weinstein
Elizabeth L. Stameshkin
**COOLEY LLP**
3175 Hanover Street
Palo Alto, CA 94304
Telephone: (650) 843-5000
Facsimile: (650) 849-7400
Email: mweinstein@cooley.com
lstameshkin@cooley.com

Kathleen R. Hartnett
Judd D. Lauter
**COOLEY LLP**
3 Embarcadero Center, 20th Floor
San Francisco, CA 94111
Telephone: (415) 693-2071
Facsimile: (415) 693-2222
Email: khartnett@cooley.com
jlauter@cooley.com

Phillip Morton
**COOLEY LLP**
1299 Pennsylvania Avenue, NW, Suite 700
Washington, DC 20004
Telephone: (202) 842-7800
Facsimile: (202) 842-7899
Email: pmorton@cooley.com

Angela L. Dunning
**CLEARY GOTTLIEB STEEN & HAMILTON LLP**
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304

By: /s/ *Maxwell V. Pritt*

**BOIES SCHILLER FLEXNER LLP**
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

Maxwell V. Pritt (SBN 253155)
Joshua M. Stein (SBN 298856)
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(415) 293-6800
mpritt@bsfllp.com
jstein@bsfllp.com

Jesse Panuccio (*pro hac vice*)
1401 New York Ave, NW
Washington, DC 20005
(202) 237-2727
jpanuccio@bsfllp.com

Joshua I. Schiller (SBN 330653)
David L. Simons (*pro hac vice*)
55 Hudson Yards, 20th Floor
New York, NY 10001
(914) 749-8200
dsimons@bsfllp.com
jischiller@bsfllp.com

*Interim Lead Counsel for Plaintiffs*

Telephone: (650) 815-4121
Facsimile: (650) 849-7400
Email: adunning@cgsh.com

*Attorneys for Defendant Meta Platforms, Inc.*

**ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(h)**

I hereby attest that I obtained concurrence in the filing of this document from each of the other signatories. I declare under penalty of perjury that the foregoing is true and correct.

Dated: December 4, 2024

BOIES SCHILLER FLEXNER LLP

*/s/ Maxwell V. Pritt*
Maxwell V. Pritt
Reed Forbush
Jay Schuffenhauer

*Attorneys for Plaintiffs*